Part 3- Session Presentations for the EPA 23rd Annual National Conference on Managing Environmental Quality Systems

April 13-16, 2004 Tampa, Florida

Analytical Software for Environmental Analysis

Review of Asymmetric Confidence Intervals and Evaluation of "WesVar" Software for Analysis of Data from NHANES Complex Surveys (Hans Allender, U.S. EPA)

Cost Effective "Collaborative Sampling" in Visual Sample Plan (VSP) Software To Estimate Means and Test Hypotheses (Richard O. Gilbert, Pacific Northwest National Laboratory)

The Use of F/S plus Geostatistical Module (John Bing-Canar, U.S. EPA)

Innovative Statistical Methodologies

Determining Detection Limits for Environmental Analyses (Thomas Georgian, U.S. Army)

A Statistical Methodology for Estimating Background Concentrations (Basil Coutant, Battelle)

Overdispersion Models for the Violation of Nitrate Concentration Limits (Nagaraj Neerchal, University of Maryland)

Assessing Environmental Health Statistics

Measuring Pesticides, Lead, Allergens, and other Dangers in Homes (John Rogers, Westat)

Pesticide Epidemiology, Biomonitoring, and Risk Assessment: Four Case Examples (Ruth Allen, U.S. EPA)

Combining Environmental Indicators (Bimal Sinha, University of Maryland)

Environmental Statistical Modeling

Determining Detection Limits for Environmental Analyses

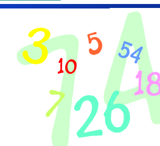
Thomas Georgian, Ph.D.

U.S. Army Corps Of Engineers
Hazardous, Toxic, and Radioactive Waste
Center Of Expertise

Method Detection Limit (MDL)

- Environmental testing industry's standard for measuring detection capability.
- Defined in 40 CFR, Part 136, Appendix B.
- Objective: To minimize "false positives" (i.e., the reporting of a "detection" when the analyte is absent) at 99% level of confidence.

Mathematical Definition of MDL



- MDL = $t_{1-p,\,\nu}\ s$
  - s = standard deviation of n ≥ 7 replicate measurements of samples spiked 1 - 10 times estimated MDL and processed through entire analytical procedure.
  - $t_{1-p,\,\nu}$ = $(1 - p)100^{th}$ percentile of Student's t distribution with ν (degrees of freedom) = n -1.
- MDL ≈ 3 s (n = 7 and p = 0.01, $t_{1-p,\,\nu} = 3.14$)

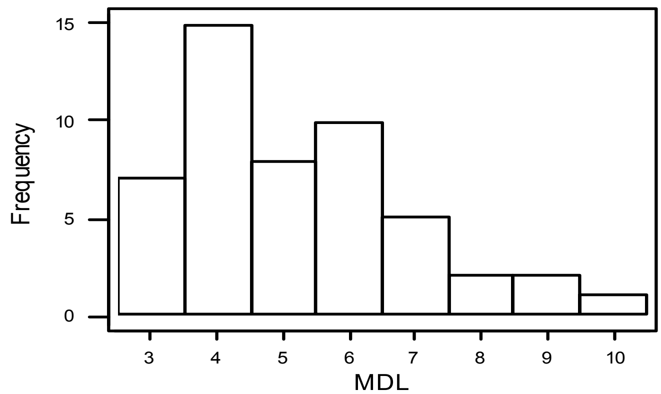
Problems with MDL: False positives

- Not conservative for limiting false positives.
  - Replicates often processed within same batch, under estimating long-term variability (i.e., s and MDL).
  - s is estimated from small set of replicates, but its uncertainty is not taken into account.
  - MDL is a "prediction limit;" minimizes false positives for only one future sample. (Probability of false positive for batch of 20 samples = $1 - (0.99)^{20} \approx 0.2$).

"MDL Histogram"

50 random sets of 7 replicates from normal distribution with μ (mean) = 10 and σ (standard deviation) = 2.

Minimum = 3, Maximum = 10, "True DL" = 5



Problems with MDL: Bias

- MDL is a function of precision (*s*) and does not take bias into account.
  - Positive Bias - Mean concentration of persistent low-level contamination can exceed MDL.
  - Negative Bias - Mean recovered analyte concentration can be less than MDL when spiked concentration is greater than MDL.

Example:

Spike = 10 ppb, MDL ≈ 2 ppb, Mean recovery ≈ 20%

Value (ppb)

1.0
1.5
2.4
1.7
1.0
2.5
2.3

x 20%

Sample Conc. = 4 ppb →

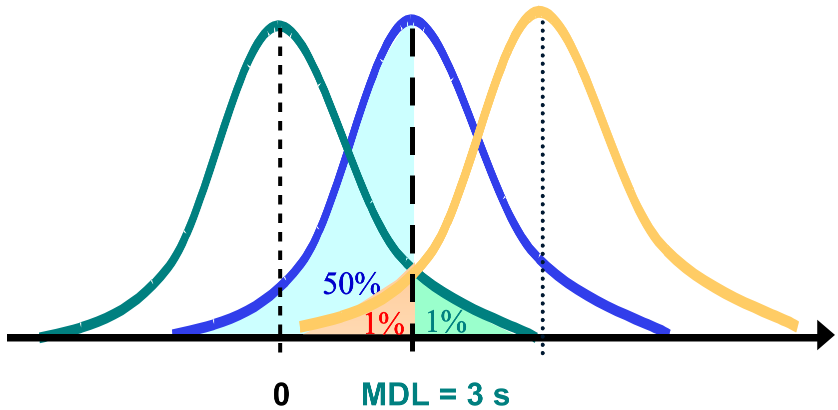
0.8 ppb < MDL = 2 ppb

We will not detect 4 ppb !

Problems with MDL: False Negatives

- MDL does not minimize false negatives !
  - Cannot reliably report "non-detects" as "< MDL."
  - False negative error at MDL is 50% (e.g., assuming a normal distribution).

False Negatives at MDL

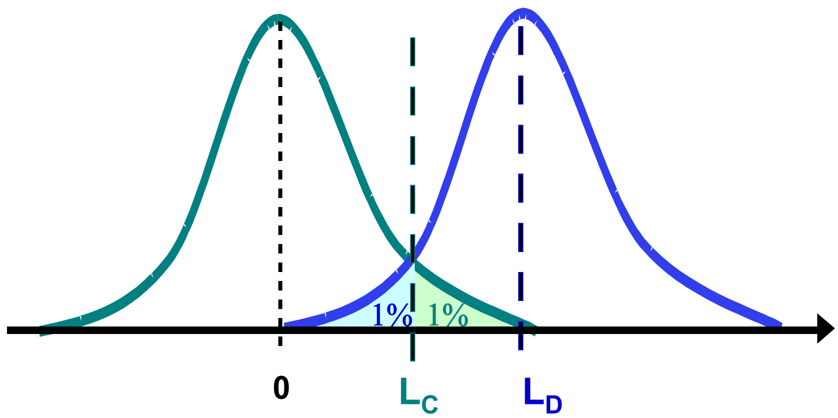


Proposed Approach

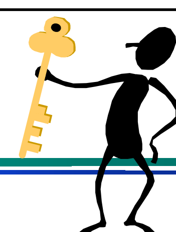
- Use two "types" of limits to measure detection capability.
  - "**Critical value**," L_C = Limit that minimizes false positives. Values greater than L_C reported as "detects." (The MDL is a type of L_C limit.)
  - "**Detection limit**," L_D = Limit that minimizes false negatives (e.g., "non-detects" reported as < L_D). (No provision for L_D limit in MDL definition.)

L_C and L_D

- L_C = Conc. greater than zero with 99% confidence.
- L_D = Conc. that will be detected with 99% confidence.

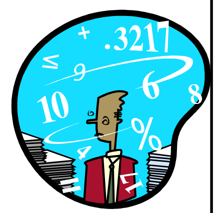


Define Conservative L_C



- Let L_C take uncertainty of *s* into account.
- L_c = Limit at which one can be $(1 - \gamma)100\%$ (e.g., 99%) confident that at least $(1 - p)100\%$ (e.g., 99%) of <u>all</u> future measurements will be less than L_C when true concentration is zero.
  - "$(1 - \gamma)100\%$ tolerance interval that covers at least $(1 - p)100\%$ of the population."

Mathematical Definition of L_C (Normal distribution, $\mu = 0$)



- $L_c = K_{1-\gamma,\, 1-p,\, \nu}\, s = z_{1-p}\,(\nu / \chi^2_{\nu,\gamma})^{1/2}\, s$
  - z_{1-p} = $(1 - p)100^{th}$ percentile of standard normal distribution ($z_{0.99} = 2.33$)
  - $\chi^2_{\nu,\gamma}$ = $\gamma\ 100^{th}$ percentile of χ^2 distribution with $\nu = n - 1$ ($\chi^2_{6,\, 0.01} = 0.872$)
  - $(1-\gamma)100\%$ UCL of s = $(\nu / \chi^2_{\nu,\gamma})^{1/2}\, s$
- $L_C \approx 6.10\, s \approx 2\, MDL$ ($p = 0.01$, $\gamma = 0.01$, $n = 7$)

L_D & False Negative Quality Control Sample (FNQS)

- $L_D \approx 2\, L_C$
- Use FNQS to verify L_C and establish L_D.
- Spike sample at $L_D \approx 2\, L_C$ (FNQS) and process through entire method. If result $X > L_C$ (and all qualitative method-specified identification criteria are fulfilled), L_C is verified and non-detects may be reported as "$< L_D$."

FNQS
(Continued)

- If FNQS not detected (i.e., $X < L_C$), increase spike (FNQS) until measured value $X > L_C$. (Calculated value of L_C not valid.) L_D = smallest FNQS that consistently gives detects and $L_C \approx L_D / 2$.
- If non-detect of FNQS due to <u>low bias</u>, then L_D = smallest FNQS that consistently results in detection. (Calculated value of L_C is retained.)
- Once L_C is determined, analyze FNQS periodically in lieu of annual MDL studies.

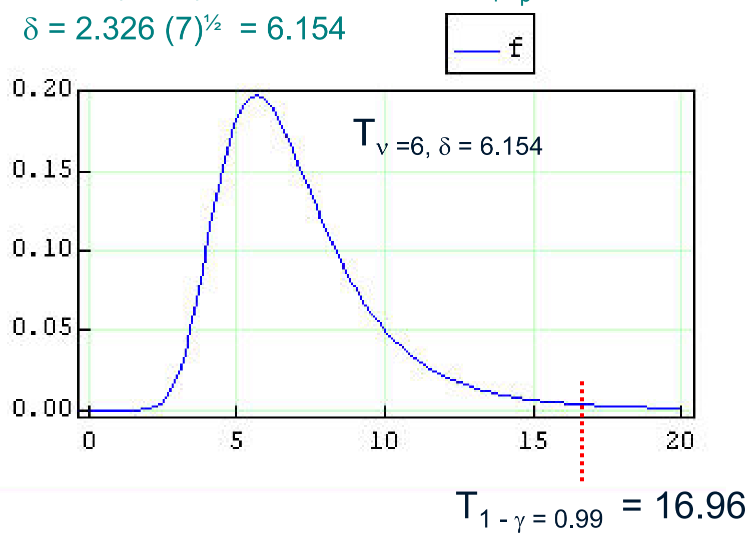
Definition of L_C for $\mu > 0$ (positive bias)

- $L_c = \langle X \rangle + K'_{1-\gamma,\,1-p,\,\nu}\, s = \langle X \rangle + T_{1-\gamma,\,\nu,\,\delta}\,(s / n^{1/2})$
  - Reduces to $L_C = K_{1-\gamma,\,1-p,\,\nu}\, s$ (slide 13) when $\mu = 0$.
  - At least (1 - p)100% (e.g., 99%) of population will be less than L_C with $(1 - \gamma)100\%$ (e.g., 99%) confidence.
  - $\langle X \rangle$ = Mean of set of $n \geq 7$ replicate method blanks
  - $T_{1-\gamma,\,\nu,\,\delta} = (1 - \gamma)100^{th}$ percentile of noncentral Student's t-distribution with $\nu = n - 1$ and noncentrality parameter $\delta = z_{1-p}\, n^{1/2}$.

Noncentral t Distribution

$n = 7$ $(\nu = 6)$, $1 - p = 0.99$, $\delta = z_{1-p}\, n^{1/2}$

$\delta = 2.326\,(7)^{1/2} = 6.154$



Determination of K′

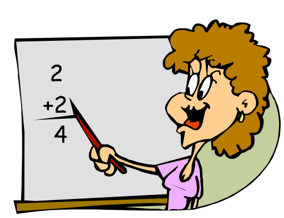
- $K'_{0.99,\, 0.99,\, 6} = T_{1-\gamma,\, \nu,\, \delta} / n^{1/2} = 16.96 / 7^{1/2} = 6.41$
- $K'_{1-\gamma,\, 1-p,\, \nu} \approx \{ z_{1-p} + [z_{1-p}^2 - a\, b]^{1/2} \} / a$

$a = 1 - (z_{1-\gamma}^2 / 2\nu)$

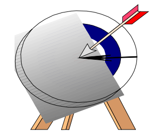
$b = z_{1-p}^2 - (z_{1-\gamma}^2 / n)$ (n ≥ 15 recommended)

$K'_{0.99,\, 0.99,\, 6} \approx 7.32$

$L_c = \langle X \rangle + K'_{0.99,\, 0.99,\, 6}\; s$



Proposed Versus MDL Approach



- Both approaches determine L_C using at least 7 replicate "clean" sample matrices.
- Both assume normality.
- Proposed approach:
  - ✓ **Takes bias into account.**
  - ✓ **More effectively minimizes false positives.**
  - ✓ **Takes false negatives into account.**
  - ✓ **Uses empirical FNQS results to check L_C and L_D**
  - ✓ **Is more cost effective.**

A Statistical Methodology for Estimating Background Concentrations

Presented at:
EPA 23rd Annual Conference on Managing Environmental Quality Systems
April 13-16, 2004
Tampa, Florida

STATISTICAL METHODOLOGIES
Thursday, April 15, 2004 - 11:00 pm

Presented by: Basil Coutant
Battelle, 505 King Avenue, Columbus, OH 43201
coutantb@battelle.org

Acknowledgements

• Joe Touma, EPA/OAQPS
• Steve Bortnick, Battelle
• Brian Biddle, Battelle
• Kristen Swinton, Battelle

Work supported by U.S. EPA/OAQPS
EPA Contract No. 68-D-02-061

Overview

- Goal and modeling perspective
- Benzene example
- Model development and advantages
- Mathematical details
- Results
- Final comments & examples

Goal

- The goal for this project was to estimate annual (or typical) background concentrations for ambient concentration measurements.
- By definition, the approach should not seek to identify what occurs during exceptional events. Rather, the desired approach needs to identify the typical background for a site.

Modeling Perspective

- The model acknowledges that evidence of background concentration levels is likely contained within all measurements, and this information should be exploited to estimate the background.

- To do this, the model treats the ambient concentrations as having two components: a fixed background and a non-negative random component from source activity.

Model Fitting

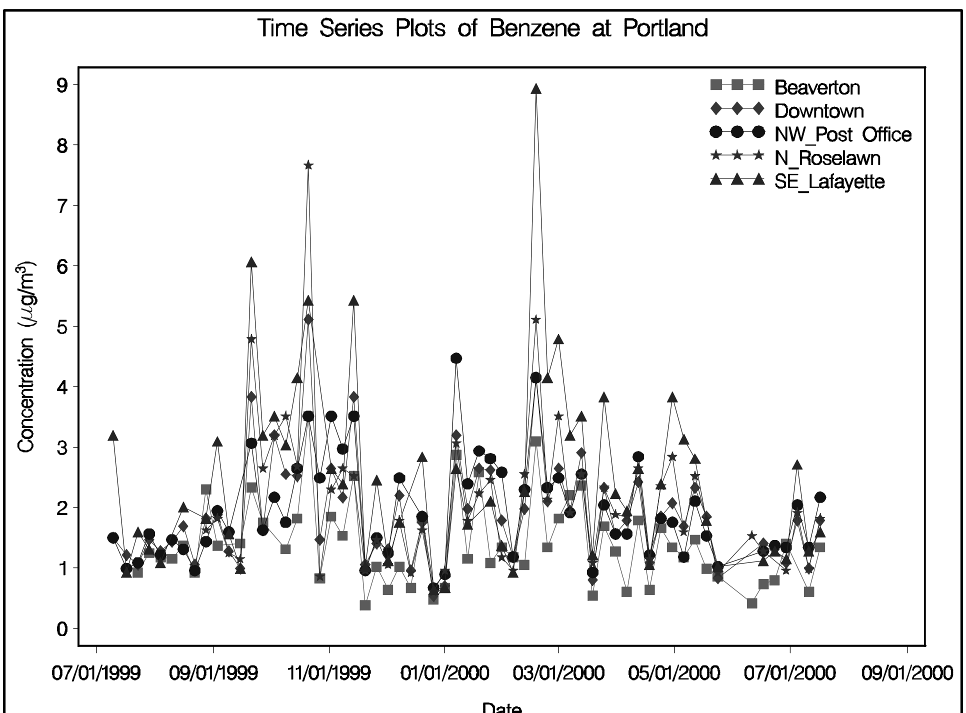
- The model was fit using SAS's NLMixed procedure that can do maximum likelihood estimation for fairly arbitrary likelihoods.

- The likelihood developed here is derived from a Gamma distribution.
  - The distribution is "shifted" from 0 to reflect the background concentration.
  - The distribution is modified in order to maintain the usual regularity conditions.

Example

- The example is from benzene concentrations observed at five monitoring sites in Portland, Oregon, from July 1999 through July 2000.
  - There are almost no data below 0.3 μg/m³,
  - about half of the data are between 0.3 and 2 μg/m³, and
  - the remainder of the data are spread out progressively thinner from 2 to 9 μg/m³.
- The minimum detection limit (MDL) is 0.1 μg/m³.
- There are high and low concentration periods, but there is no obvious seasonal trend in the Portland data.

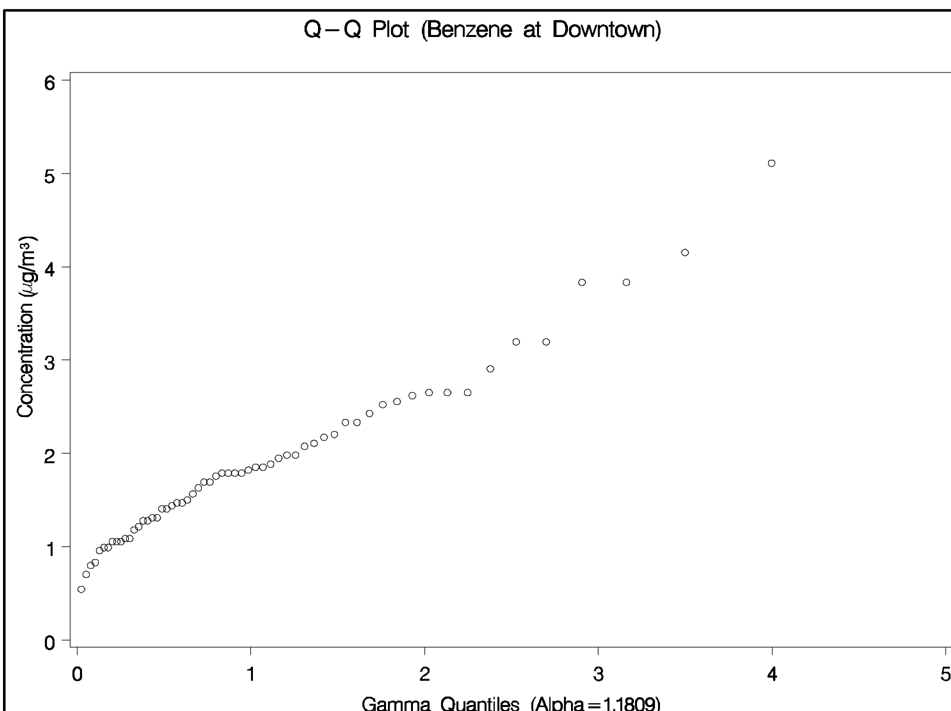
April 13-16, 2004 EPA Conf. - Tampa, FL 7



Check on the Choice of the Gamma

- Given that Log-normal distributions are the "default" for non-negative environmental data. Why use a Gamma distribution?

  1. The PDF does not have a factor of "x" in the denominator for the range of parameters considered. As a result, the likelihood is a better behaved numerically.
  2. It fits well.



Development of the Likelihood

- The goal is to capture the mean long-term vertical shift rather than the minimum shift.
  - The model was modified from a shifted gamma to a model that treats values near and below the mean shift differently from the remainder of the data.
  - The data near [within 2 times the minimum detection limit (MDL)] or below the shift are treated as random noise.
  - The adjustment introduced to make this modification has several advantages.

Advantages of the Modified, Shifted Gamma

- It will naturally handle below MDL data without additional modifications.
  - Moreover, it is modified in a way that does not require one to know the relative sizes of the MDL and the shift.
- The model is continuous and always supported on $[0, \infty)$.
- This results in a model that satisfies regularity conditions [1] so that the standard errors can be estimated via large sample theory.

The Model

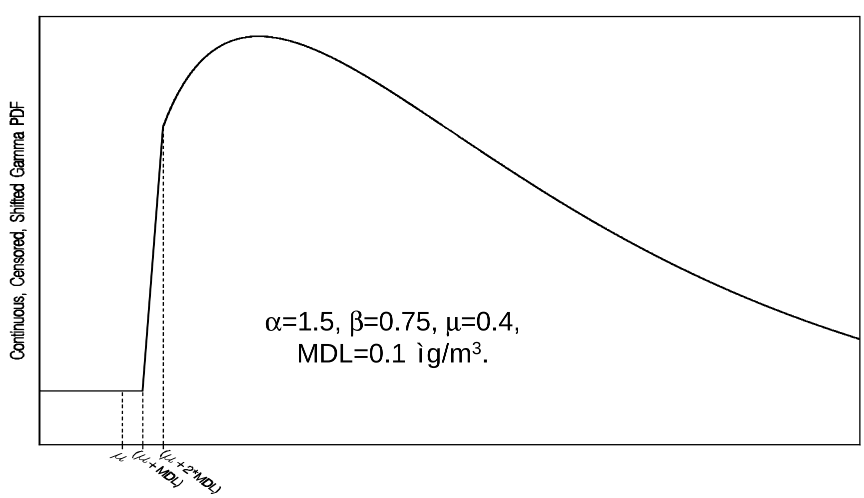
L(ì, á, â) = "shifted gamma" (i.e., y_i = ì + $å_i$).

ì = the constant background, $å_i$ ~ Gamma

$$L(m, a, b) = \begin{cases} \text{"shifted gamma"} & \textit{if } y_i \geq m + 2 \cdot MDL \\ \text{"constant"} & \text{otherwise} \end{cases}$$

$$L(m, a, b) = \begin{cases} \text{"shifted gamma"} & \textit{if } y_i \geq m + 2 \cdot MDL \\ \text{"linear"} & \textit{if } m + MDL \leq y_i < m + 2 \cdot MDL \\ \text{"constant"} & \textit{otherwise} \end{cases}$$

Typical PDF



Mathematical Details

Forcing the likelihood to be continuous causes the likelihood to satisfy a regularity condition, namely that one can differentiate under the integral. To see why this works, consider a hypothetical case where *f* is a continuous pdf with the form:

$$f(x,g) = \begin{cases} g(x,g) & if \;\; x \le r(g) \\ h(x,g) & if \;\; r(g) < x \end{cases}$$

Regularity Conditions

Then for a continuous s(x) with finite expectation:

$$\begin{aligned} \frac{d}{dg}\int_{-\infty}^{\infty} s(x)\cdot f(x,g)dx &= \frac{d}{dg}\int_{-\infty}^{r(g)} s(x)\cdot g(x,g)dx + \frac{d}{dg}\int_{r(g)}^{\infty} s(x)\cdot h(x,g)dx \\ &= \int_{-\infty}^{r(g)} s(x)\cdot\frac{\partial}{\partial g} f(x,g)dx + s(r(g))g(r(g),g)r'(g) + \\ &\quad \int_{r(g)}^{\infty} s(x)\cdot\frac{\partial}{\partial g} f(x,g)dx - s(r(g))h(r(g),g)r'(g) \\ &= \int_{-\infty}^{\infty} s(x)\cdot\frac{\partial}{\partial g} f(x,g)dx \end{aligned}$$

Finding the Constant

$$1 = \int_{0}^{m+MDL} k\,dt + \int_{m+MDL}^{m+2\cdot MDL} line\ dt + \int_{m+2\cdot MDL}^{\infty} Shifted\ Gamma\ pdf\ dt \quad \text{or}$$

$$1 = k \cdot (m + MDL) + \frac{1}{2}(k + y) \cdot MDL + (1 - CDF(2 \cdot MDL))$$

where k is the value of the constant, y is the value of the shifted gamma PDF at (μ+2*MDL) = the standard gamma PDF at 2MDL, and the CDF(2MDL) is the value of the standard Gamma CDF at 2MDL.

$$k = \frac{2 \cdot CDF(2 \cdot MDL) - y \cdot MDL}{(2 \cdot m + 3 \cdot MDL)}$$

Results

Site	Sample Size	Mean	Stand. Dev.	Max	Min	Background Estimate	Standard Error
Beaverton	56	1.3840	0.6948	3.5127	0.3832	0.4067	0.2034
Downtown	60	1.8953	0.8908	5.1094	0.5429	0.5491	0.0593
NW Post Office	59	1.9204	0.8738	4.4707	0.1000	0.7359	0.0563
N Roselawn	52	2.0972	1.2321	7.6641	0.6067	0.7127	0.0644
SE Lafayette	55	2.4844	1.5601	8.9415	0.6387	0.7364	0.0784
All Sites	282	1.9511	1.1342	8.9415	0.1000	0.6282	0.0483

Final Comments

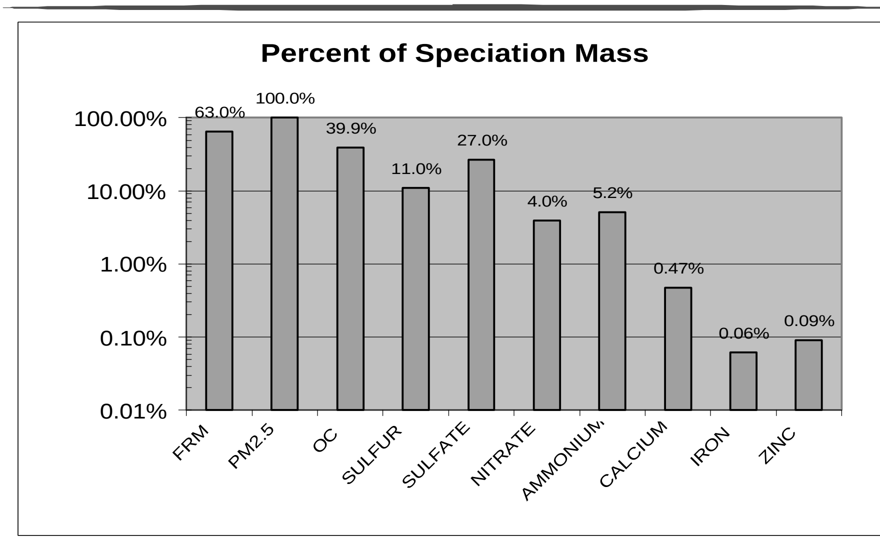
- The results are in good agreement with the intuition from the data and are comparable to the 1996 NATA background estimate of 0.48 $\mu g/m^3$.
- Note that the Beaverton site is located in an adjacent, more rural county and is separated from the other sites by a low ridge of hills.
- The method was applied to the data in the Air Toxics Archive to get background estimates for a wide variety of HAPs from 1995-2002 data.

Another Example?

- Cleveland, OH, STN data
  - Organic carbon is not blank corrected and there is reason to suspect a contamination problem.
  - Applying the background estimation algorithm to the STN data from Cleveland, OH, results in most species with an estimated “background” that is less than 2 times their MDL.
  - The exceptions are:

FRM mass	PM2.5 mass	Organic Carbon	
Ammonium	Sulfur	Sulfate	Nitrate
Calcium	Iron	Zinc	

Profile Plot



What is it?

- The estimate for the speciation sampler's $PM_{2.5}$ mass is 3.3 (0.7) ìg/m^3.
- The estimate for the FRM mass is 2.1 (0.8) ìg/m^3.
- 3 times sulfur ~ sulfate (as expected).
- The reconstructed mass is consistent with the "measured" mass.
- Not enough ammonia.
- If the Ca & Fe are from dirt, then where is the Al?

Region 5 Changes In Estimated Hazard Exposure and Demographic Characteristics: 1990-2000

Lawrence Lehrman and Arthur N. Lubin

Introduction

- Procedures used to show areas of migration of population and changes in toxicity hazard estimates between 1990 and 2000.
- Changes in hazard estimates were compared with changes in population density.

Data-base Development

- Developed data base to combine EPA approved risk coefficient values (TRI RSEI) with Census block group level information 1990 - 2000.

Risk Screening Environmental Indicators (RSEI)

- **RSEI was produced by the Economics, Exposure, and Technology Division Office of Pollution Prevention and Toxics.**
- **Risk-related score is a unitless value proportional to the potential risk-related impact of each element.**
- **Scores per TRI facility are derived by summing the estimated risks from each element**
- **RSEI was used to compare relative risk levels over time.**

U.S. Census block-group level data

- Block-group level data were used is that it is the highest resolution geographic area that incorporates sampled data.
- Sample data includes parameters of population density, new homes (built since the last census) , minority populations and low income.

Development of Region 5 Grid DataBase

- Grid cells are spatially uniform
- Grid cells are excellent for incorporating polygon features with inconsistent boundaries.
- Product allows us to complete boolean queries, algebraic functions, and statistical analysis from one table.

Grid Cell Resolution and Data Allocation techniques

- Cell size 3x3 Km consistent with the resolution of block groups in suburban expected to be experiencing growth.
- Small spatial variations in the demographics data smoothed by allocating demographics based on centroid buffered to 5k.
- Populations living near edge of the cell to be considered as having an influence.

Grid Cell Hazard Allocation Technique

- Centroid of grid cell buffered to 5k and intersected with RSEI scores.
- Releases from facilities near the edge of the cell to be considered as having an influence.
- Region 5 grid consists of 99,639 3 by 3 km grid cells.

Attributes for Analysis

- **Attributes are the 1990 and 2000 Census Demographics (population density, new homes built since last census, and the Environmental Justice parameters of minority, low income and poverty levels) and the 1990 and 2000 RSEI hazard densities.**
- **Population and toxicity changes over time were calculated by subtracting 1990 from 2000 results.**

Quartiling Changes

- Population and hazard changes were quartiled and the results mapped.
- (map images here)May be not here

Analysis and Results

- Determine areas of apparent increasing risk using the RSEI Hazard numbers and determine whether changes differ in areas with distinct demographic attributes.
- Initially calculated averages for the demographic change variable (2000-1990) and the 2000 risk estimate.

Analysis and Results Ctd.

- Quartile means of the 2000 hazard estimate indicate that persons tend not to migrate to areas with higher exposures.
- However, the demographic characteristics of each of the population growth attributes are relatively similar.

Table 1 here

Place figure 1 here

Analysis and Results ctd.

- Figure 1 shows the apparent relationships between the population change quartiles and the hazard 2000 changes .

Analysis and Results ctd.

- **The Figure also displays the areas where the following conditions are present - both population change and hazard growth estimates are in the upper or lower quartile or half. Areas not colored per the legend are where population growth and hazard estimates are in opposing quartiles or halves.**

Analysis and Results ctd.

- Because a high proportion of the Chicago PMSA is not colored per the legend, Figure 1 further suggests an aversion to migrating into areas with higher hazard estimates. Further, there are substantial areas where the relative risk estimates are decreasing; especially in areas with reduced population growth.

place Figures 2 and 3 here

Analysis and Results ctd.

- Figures 2 and 3 show areas where there are decreasing/increasing RSEI numbers for Region 5 and the Chicago PMSA respectively.
- The maps were developed by quartiling the cells hazard estimates for 2000 and 1990 and subtracting the 1990 from the 2000 quartile. Positive results show where relative risk has increased and vice versa.

Analysis and Results ctd.

- In Figures 2 and 3 there are high proportions of the cells which had lower relative risk estimates in 2000 versus 1990. The maps must be interpreted cautiously due to several areas having artificially increased risk levels due to the expanded number of TRI chemicals, new industries in the TRI and reduced reporting thresholds for several chemicals.

Place Figures 4 and 5 here

Analysis and Results ctd.

- Another interest was whether or not the population change quartiles actually yield relatively homogeneous subareas relative to the demographics and the estimated 2000 hazard toxicity exposure. This was tested using discriminant analysis.

Analysis and Results ctd.

- Discriminant analysis uses the generalized squared distances among the block groups to categorize the groups into categories while considering the original quartile categorization probabilities using the pooled covariance matrix. Version 8.0 of SAS was used for the calculations.
- Figures 4 and 5 show the original and statistical groupings respectively.

Analysis and Results ctd.

- The discriminant analysis and original quartiles for the entire Region had a fairly high level of categorization agreement (about 65 per cent). Furthermore, it was relatively unusual for discriminant analysis to yield reallocations of more than one category different from the original quartiles.

Analysis and Results ctd.

- Thus, the relatively simplistic quartiling approach seems to provide groupings that are not woefully inadequate for creating at least somewhat homogeneous groupings.
- Relatively homogeneous subareas are potentially valuable for a wide range of applications including developing more efficient sampling plans and the targeting of environmental mitigation efforts.

Summary and Results

- The basic results are summarized as follows:
  - **Areas in Region 5 with greater risk estimates tend to have lower population growth. However, risk estimates and changes in demographics seem to be minimally related.**
  - **There are several areas where the relative risk is decreasing; especially in those areas with lower population growth.**

Summary and Results ctd.

- **The analysis demonstrated that a substantial proportion of the areas investigated had lower relative risk estimates in 2000 versus 1990.**
- **The discriminant analysis demonstrated that the relatively simplistic simplistic approach of quartiling based on population growth provides at least somewhat homogeneous area groupings in terms of 2000 risk estimates and the selected demographic characteristics.**

Recommended Future Efforts

- Present effort should definitely be viewed as a work in progress. There are several areas where future efforts should be directed:
  - **The investigation could be focused on additional areas (smaller or larger) or larger to determine if the results would be similar in other locales.**

Recommended Future Efforts ctd.

- Specific characteristics which account for the relative consistency of the groupings are uncertain.
- Similar investigations are encouraged to determine the findings which would emerge using a particular type of pollutant and/or facility.

Recommended Future Efforts ctd.

- The present data base development and data analysis relied upon the combined usage of several software packages and data bases. Could similar procedures be done using additional software packages and/or data bases to expand to a multimedia effort?

Recommended Future Efforts ctd.

- **The previous suggestions are only a few of the above are only some of the future efforts which could be done. Any suggestions for potential future future directions and/or methods would be appreciated.**